VŠB – Technická univerzita Ostrava
Fakulta elektrotechniky a informatiky
Katedra aplikované matematiky

Zvláštnosti pri použití testu ANOVA

Peculiarities by the use of ANOVA test

2019

Andrea Jendrisková

VŠB - Technická univerzita Ostrava
Fakulta elektrotechniky a informatiky
Katedra aplikované matematiky

Zadání bakalářské práce

Student: **Andrea Jendrisková**

Studijní program: B2647 Informační a komunikační technologie

Studijní obor: 1103R031 Výpočetní matematika

Téma: Zvláštnosti při použití tesu ANOVA
Peculiarities by the use of ANOVA test

Jazyk vypracování: čeština

Zásady pro vypracování:

1. Studium teoretických základů k testu ANOVA
2. Srovnání testu ANOVA s odpovídajícími párovými testy.
3. Analýza výběrového mezitřídního rozptylu.
4. ANOVA při porušení předpokladů.
5. Demonstrace zvláštností na kontrolních příkladech, závěry.

Seznam doporučené odborné literatury:

* Briš R., Litschmannová M.,STATISTIKA I. pro kombinované a distanční studium, Elektronické skriptum VŠB TU Ostrava,2004.
* Weerahandi S.: Exact Statistical Methods for Data Analysis, Springer-Verlag New York, Inc., 1995, ISBN 0-387-94360-9.
* Briš, R., Litschmannová M.: STATISTIKA II., 2007 Ostrava:VŠB – Technická univerzita Ostrava, 2007, ISBN 978-80-248-1482-7, elektronické skriptum.

Formální náležitosti a rozsah bakalářské práce stanoví pokyny pro vypracování zveřejněné na webových stránkách fakulty.

Vedoucí bakalářské práce: **prof. Ing. Radim Briš, CSc.**

Datum zadání: 01.09.2018

Datum odevzdání: 30.04.2019

VYSOKÁ ŠKOLA BÁŇSKÁ-TECHNICKÁ UNIVERZITA OSTRAVA
Fakulta elektrotechniky a informatiky
1D16

prof. RNDr. Jiří Bouchala, Ph.D.
vedoucí katedry

prof. Ing. Pavel Brandštetter, CSc.
děkan fakulty

Prehlasujem, že som túto bakalársku prácu vypracovala samostatne. Uviedla som všetky literárne pramene a publikácie, z ktorých som čerpala.

V Ostrave 30. apríla 2019

....

Rada by som poďakovala vedúcemu práce prof. Ing. Radimovi Brišovi, CSc. za čas, ktorý mi venoval, konzultácie, poskytnuté podklady a za odborné vedenie pri tvorbe bakalárskej práce, pretože bez neho by táto práca nevznikla.

Abstrakt

Táto práca popisuje parametrický štatistický test ANOVA s ohľadom na jeho predpoklady a zvláštnosti s ním spojené. Prvá kapitola popisuje základné definície a metódy používané pri analýze rozptylu. V druhej kapitole sa zameriavame na predpoklady samotného testu a ďalšie postupy pri ich nesplnení. Posledná kapitola je venovaná porovnaniu dvojvýberového t-testu a testu ANOVA, konkrétne v prípade dvoch nezávislých výberov. Kapitoly obsahujú odpovedajúce ilustračné príklady. Príklady sú riešené pomocou programovacieho jazyka R za použitia softwaru RStudio.

Kľúčové slová: ANOVA, Dvojvýberový t-test, Súčet štvorcov, Rozptyl, F-pomer, p-hodnota, Nulová hypotéza

Abstract

This thesis describes parametric statistical test ANOVA, its assumptions and peculiarities. The first chapter describes basic definitions and methods used by analysis of variance. The main focus in the second chapter are assumptions of the ANOVA test and additional steps, if the assumptions are not met. The last chapter describes comparison between two-sample t-test and ANOVA in case of two independent samples. Chapters cointain demonstrations of concrete problems. The solution of the problems is done via programming language R using software RStudio.

Key Words: ANOVA, Two-sample t-test, Sum of Squares, Variance, F-ratio, p-value, Null Hypothesis

Obsah

Zoznam obrázkov

Zoznam tabuliek

1 Úvod

V tejto práci sa venujeme analýze rozptylu. Analýza rozptylu, alebo ANOVA z angl. (**AN**alysis **O**f **VA**riance) je parametrický test o zhodne stredných hodnôt jednotlivých výberov z populácie. Testuje, či na hodnotu náhodnej veličiny štatisticky významne vplýva hodnota niektorého znaku konečného počtu variant (minimálne dvoch), ktorý pozorujeme. Pomocou variant znaku rozdeľujeme testované subjekty do skupín.

Vlastnosti a definície sú uvedené v kapitole číslo 2, kde sa bližšie venujeme aj vlastnostiam výberových rozptylov. Výberové rozptyly ďalej využívame pri zostavení F-pomeru.

Analýza rozptylu vo svojej parametrickej forme predpokladá nezávislosť výberu, normalitu rozdelenia, homoskedasticitu, inak aj zhodnosť rozptylov jednotlivých výberov. V prípade, že niektorý z týchto predpokladov nie je splnený, výsledok testu ANOVA nemôžeme považovať za korektný a musíme použiť niektorý z iných testov, o ktorých viac píšeme v kapitole 3. V bakalárskej práci sú tieto testy rozoberané len veľmi stručne a definície sú prebraté z iných literatúr, keďže hlavným cieľom tejto práce je samotná ANOVA a jej vlastnosti.

V kapitole 4 sa venujeme porovnaniu medzi dvojvýberovým t-testom a analýzou rozptylu. O tomto porovnavaní má zmysel uvažovať, pretože vieme, že ANOVA je rozšírením testov o stredných hodnotách dvoch nezávislých výberov s normálnym rozdelením.

Jednotlivé kapitoly obsahujú príklady na ktorých demonštrujeme postup pri teste ANOVA. Pri zamietnutí nulovej hypotézy postupujeme post-hoc analýzou. V tejto práci budeme používať iba LSD metódu.

Pri vypočítavaní výsledných pozorovaných hodnôt a vykresľovaní grafov budeme používať software RStudio.

2 Základne definície a úvod do jednofaktorovej analýzy rozptylu

ANOVA alebo analýza rozptylu nám umožnuje porovnávať stredné hodnoty dvoch a viac nezávislých náhodných výberov. Náhodný výber je taký, v ktorom má každá jednotka populácie známu, avšak nenulovú, pravdepodobnosť zaradiť sa do výberu.

ANOVA patrí medzi parametrické štatistické testy. Analýza rozptylu skúma závislosť intervalovej či pomerovej premennej X na nominálnej premennej A, ktorá má aspoň dve varianty. Premenná A sa nazýva faktor a jej varianty sa nazývajú úrovne faktoru.

Závisloť X na A sa prejaví tým, že existuje štatisticky významný rozdiel v stredných hodnotách premennej X v spomínaných náhodných výberoch, ktoré vznikli triedením podľa variant premennej A.

V tejto kapitole si zavedieme základné pojmy, ktoré využívame pri analýze rozptylu.

2.1 Základné definície

Majme k populácii. Majme náhodný výber i s počtom pozorovaní n_i, kde *i=1,...,k*[5]:

$Výber\ 1 : (X_{11}, X_{12}, ..., X_{1j}, ..., X_{1n_1})$
$Výber\ 2 : (X_{21}, X_{22}, ..., X_{2j}, ..., X_{2n_2})$
...
$Výber\ i : (X_{i1}, X_{i2}, ..., X_{ij}, ..., X_{in_i})$
...
$Výber\ k : (X_{k1}, X_{k2}, ..., X_{kj}, ..., X_{kn_k})$

Budeme mať teda k náhodných výberov. Predpokladajme, že výbery sú nezávislé a pochádzajú z normálneho rozdelenia s rovnakým rozptylom $N(\mu_k, \sigma^2)$.

Vyššie popísaný model sa označuje model analýzy rozptylu jednoduchého triedenia. V tomto prípade pozorujeme nasledujúci matematický model:

$$X_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim\ N(0, \sigma^2), \quad i = 1, ...k, \quad j = 1, ..., n_i, \tag{1}$$

kde X_{ij} je $j-$té pozorovanie z $i-$tej skupiny, α_i je účinok faktoru A v $i-$tom výbere, μ je spoločná stredná hodnota a ϵ_{ij} sú nezávislé náhodné veličiny vyjadrujúce bližšie nešpecifikované náhodne chyby, ktoré vznikajú pri každom meraní.

Predtým, ako prejdeme k definíciam a označeniam, si sformulujeme problém. Testujeme nulovú hypotézu, ktorá hovorí, že stredné hodnoty jednotlivých výberov sa štatisticky významne nelíšia:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k, \text{ kde } \mu_i = \mu + \alpha_i \text{ pre } i = 1, ...k$$

oproti alternatívnej hypotéze, ktorá hovorí, že sa aspoň jedna dvojica stredných hodnôt štatisticky významne líši:

$$H_A : \neg H_0$$

Za predpokladu nulovej hypotézy sa model 1 zmení na:

$$X_{ij} = \mu + \epsilon_{ij}, \quad \epsilon_{ij} \sim \ N(0, \sigma^2), \quad i = 1, ...k, \quad j = 1, ..., n_i. \tag{2}$$

Ďalej budeme používať nasledujúce značenie:

- Celkový počet pozorovaní vo všetkých k výberoch:

$$N = \sum_{i=1}^{k} n_i \tag{3}$$

- Výberové priemery jednotlivých k výberov:

$$\bar{X}_1, \bar{X}_2, ..., \bar{X}_k,$$

  a teda výberový priemer i-tého náhodného výberu:

$$\bar{X}_i = \frac{\sum_{j=1}^{n_i} X_{ij}}{n_i}, i = 1, ..., k. \tag{4}$$

  Priemer zo všetkých pozorovaných hodnôt potom budeme značiť ako $\bar{X}$.

- Výberové rozptyly jednotlivých k výberov značíme ako:

$$\bar{S}_1^2, \bar{S}_2^2, ..., \bar{S}_k^2,$$

  a teda výberový rozptyl i-tého náhodného výberu:

$$\bar{S}_i^2 = \frac{\sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2}{n_i - 1}, i = 1, ..., k \tag{5}$$

- Výberovú smerodajnú odchýlku i-tého náhodného výberu definujeme ako kladnú odmocninu výberového rozptylu:

$$\bar{S}_i = \sqrt{\bar{S}_i^2} = \sqrt{\frac{\sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2}{n_i - 1}}, i = 1, ..., k \tag{6}$$

Zavedieme definície súčtu štvorcov alebo variabilít:

Definícia 2.1.1 Totálny súčet štvorcov (alebo totálnu variabilitu) definujeme ako

$$SS_T = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X})^2, \tag{7}$$

kde $\bar{X}$ je výberový priemer zo všetkých pozorovaných hodnôt.

Totálnu variabilitu závislej premennej je možné rozdeliť na dve časti, na variabilitu vnútri skupín (vnútorná variabilita) a na variabilitu medzi skupinami (medzitriedna variabilita).

Definícia 2.1.2 Vnútornú variabilitu definujeme ako

$$SS_W = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2 = \sum_{i=1}^{k}(n_i - 1)S_i^2, \tag{8}$$

Definícia 2.1.3 Medzitriednu variabilitu definujeme ako

$$SS_B = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2 \tag{9}$$

Medzi jednotlivými variabilitami potom existuje vzťah:

$$SS_T = SS_W + SS_B$$

Rozptyl je hodnota, ktorá udáva, ako veľmi sú jednotlivé dáta vzdialené od priemernej hodnoty. Väčšie hodnoty znamenajú väčšiu vzdialenosť od priemernej hodnoty, a teda väčšiu disperziu.

Definícia 2.1.4 Vnútorný výberový rozptyl definujeme ako podiel vnútornej varibility a rozdielu počtu všetkých dát a počtu nezávislých výberov:

$$S_W^2 = \frac{SS_W}{N - k} \tag{10}$$

Definícia 2.1.5 Medzitriedny výberový rozptyl definujeme ako podiel medzitriednej variability a počtu nezávislých výberov zmenšený o 1:

$$S_B^2 = \frac{SS_B}{k - 1} \tag{11}$$

Vlastnosti týchto výberových rozptylov [1]:

- Vnútorný výberový rozptyl je nestranným odhadom rozptylu nezávisle na H_0.

$$ES_W^2 = \frac{1}{N-k}E(\sum_{i=1}^{k}(n_i - 1)S_i^2) = \frac{1}{N-k}\sum_{i=1}^{k}(n_i - 1)E(S_i^2) = \sigma^2,$$

  pretože $E(S_i^2) = \sigma^2$.

- ES_B^2 je nestranným odhadom spoločného rozptylu, ak platí H_0.

$$ES_B^2 = \sigma^2 + \frac{1}{(k-1)} \sum_{i=1}^{k} n_i(E\bar{X} - E\bar{X}_i)^2$$

  potom $ES_B^2 = \sigma^2$ práve vtedy, keď platí H_0. Túto vlastnosť si dokážeme v ďalšej podkapitole.

2.1.1 Analýza medzitriedneho výberového rozptylu

Veta 2.1.1.1. $ES_B^2 = \sigma^2 + \frac{1}{(k-1)} \sum_{i=1}^{k} n_i(E\bar{X} - E\bar{X}_i)^2$, potom $ES_B^2 = \sigma^2$ práve vtedy, keď platí H_0.

Dôkaz. Využijeme definíciu 11. Najdôležitejším krokom pri dôkaze bude úprava a stredovanie medzitriednej variability SS_B:
Podľa vzorca 9 vieme, že

$$SS_B = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2$$

$$\sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2$$

môžeme rozšíriť o $(E\bar{X}_i - E\bar{X}_i)$ a $(E\bar{X} - E\bar{X})$ a správnou úpravou dostávame

$$\sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2 = \sum_{i=1}^{k} n_i((\bar{X}_i - E\bar{X}_i) - (\bar{X} - E\bar{X}) - (E\bar{X} - E\bar{X}_i))^2$$

Výraz v zátvorkách môžeme rozložiť podľa vzorca $(a-b-c)^2 = a^2+b^2+c^2-2ab+2bc-2ac$, z čoho dostávame:

$$\begin{aligned}&(\bar{X}_i - E\bar{X}_i)^2 + (\bar{X} - E\bar{X})^2 + (E\bar{X} - E\bar{X}_i)^2 - 2(\bar{X}_i - E\bar{X}_i)(\bar{X} - E\bar{X}) - \\ &- 2(\bar{X}_i - E\bar{X}_i)(E\bar{X} - E\bar{X}_i) + 2(\bar{X} - E\bar{X})(E\bar{X} - E\bar{X}_i) \qquad (12)\end{aligned}$$

Položme

$$temp = \sum_{i=1}^{k} n_i[(\bar{X} - E\bar{X})^2 - 2(\bar{X}_i - E\bar{X}_i)(\bar{X} - E\bar{X})] \qquad (13)$$

V nasledujúcich krokoch budeme upravovať vzťah 13.

$$temp = (\bar{X} - E\bar{X})^2 N - 2(\bar{X} - E\bar{X}) \sum_{i=1}^{k} n_i(\bar{X}_i - E\bar{X}_i),$$

kde $N = \sum_{i=1}^{k} n_i$.

Použitím vzorca 4 dostávame:

$$temp = (\bar{X} - E\bar{X})^2 N - 2(\bar{X} - E\bar{X}) \sum_{i=1}^{k} n_i(\frac{\sum_{j=1}^{n_i} X_{ij}}{n_i} - E\frac{\sum_{j=1}^{n_i} X_{ij}}{n_i})$$

Vykrátime n_i a zátvorku rozšíreme o N:

$$temp = (\bar{X} - E\bar{X})^2 N - 2(\bar{X} - E\bar{X})(\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}}{N} - E\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}}{N})N$$

Zjednodušíme:

$$temp = (\bar{X} - E\bar{X})^2 N - 2(\bar{X} - E\bar{X})(\bar{X} - E\bar{X})N = -N((\bar{X} - E\bar{X})^2), \tag{14}$$

pretože ak platí H_0, potom všetky $E\bar{X}_i$ sú rovnaké a rovnajú sa celkovému $E\bar{X}$.

Vzťah 14 dosadíme do rovnice číslo 12. ESS_B potom vyjadríme ako

$$ESS_B = \sum_{i=1}^{k} n_i(\frac{\sigma^2}{n_i} + (E\bar{X} - E\bar{X}_i)^2) + (-N\frac{\sigma^2}{N})$$

$$ESS_B = k\sigma^2 + \sum_{i=1}^{k} n_i(E\bar{X} - E\bar{X}_i)^2) - \sigma^2 = \sigma^2(k-1) + \sum_{i=1}^{k} n_i(E\bar{X} - E\bar{X}_i)^2 \tag{15}$$

ESS_B vydelíme $k-1$ a dostávame:

$$\frac{ESS_B}{k-1} = \sigma^2 + \frac{1}{k-1}\sum_{i=1}^{k} n_i(E\bar{X} - E\bar{X}_i)^2 \tag{16}$$

Vieme, že ak platí H_0, potom všetky $E\bar{X}_i$ sú rovnaké a rovnajú sa celkovému $E\bar{X}$. Potom:

$$\frac{ESS_B}{k-1} = \sigma^2$$

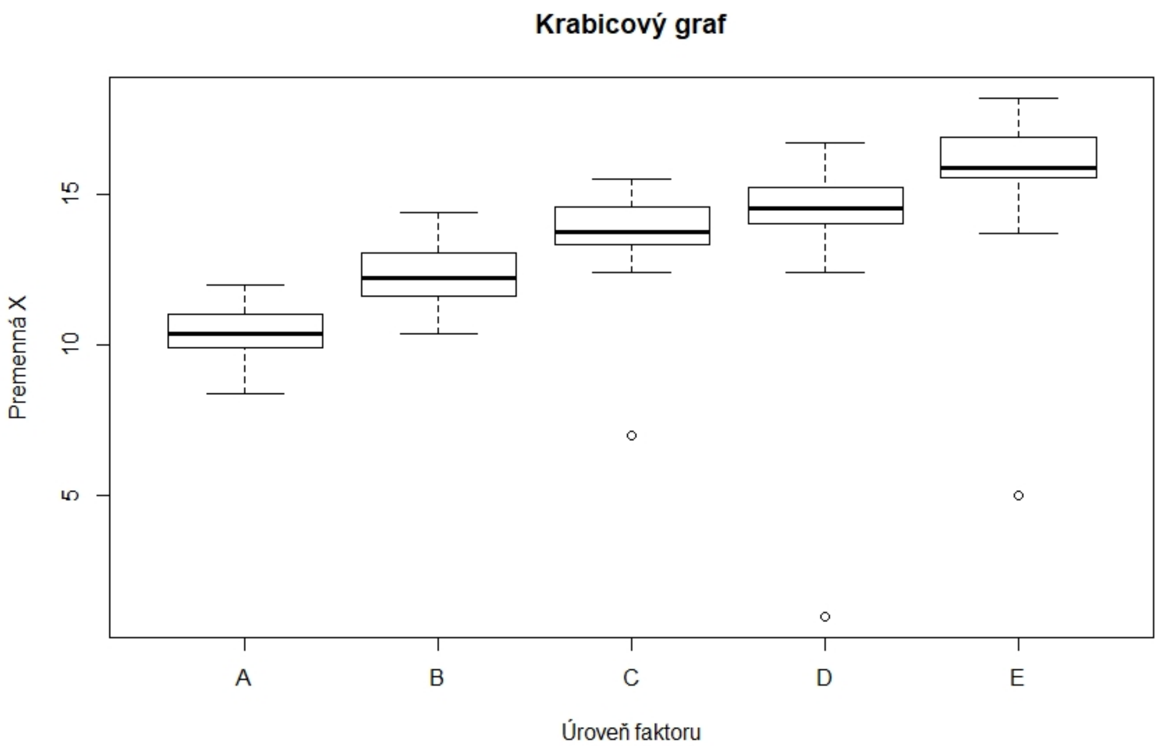
$$ES_B^2 = \sigma^2 \tag{17}$$

Dôsledok: Ak platí H_0, tak S_B^2 je nestranným odhadom spoločného rozptylu.

2.2 Exploračná analýza

Pred samotným testom je potrebná vizualizácia dát a ich analyzovanie, aby nedošlo k analyzovaniu dát nesprávnym testom a podobne. V prípade, že sú dáta dostatočne malé, môžeme použiť bodový graf, avšak pri väčšom množstve dát sa odporúča viacnásobný krabicový graf. Mimo iného môžeme vidieť aj odľahlé pozorovania, ktoré by mohli výsledok testu znehodnotiť.

V prípade, kedy poznáme dôvod vzniku odľahlých pozorovaní a predpokladáme, že už nenastanú, môžeme ich vylúčiť z ďalšieho testovania. Známe dôvody takýchto pozorovaní môžu byť napr. preukázateľné zlyhanie techniky alebo ľudí, chybné meranie, preklep, hrubá chyba a podobne. V prípade zanechania takýchto pozorovaní v dátach použijeme radšej Kruskal-Wallisov test, ktorý je neparametrickou obdobou anlýzy rozptylu a používa sa napr. pri nesplnení predpokladu normality dát.
Ukážka krabicového grafu s odľahlými hodnotami:



Obr. 1: Ukážka krabicového grafu

2.3 Testovacie kritérium F-pomer

F-pomer je pomer dvoch rozptylov. V našom prípade je to pomer medzitriedneho výberového rozptylu a vnútorného výberového rozptylu[1]:

$$F = \frac{S_B^2}{S_W^2} \tag{18}$$

V prípade, že platí H_0 (stredné hodnoty čitateľa a menovateľa F-pomeru sú rovnaké), je F-pomer číslo blízke 1. Avšak v prípade, že H_0 neplatí, je F-pomer výrazne väčšie číslo ako 1, čiže variabilita medzi skupinami bude oveľa vačšia ako vnútri skupín. V prípade zamietnutia H_0 je čitateľ v F-pomeru výrazne väčší, čo plynie z vety 2.1.1.1.

V prípade platnosti H_0 má F-pomer Fisher-Snedecorove rozdelenie s k-1 stupňami voľnosti v čitateli a s N-k stupňami voľnosti v menovateli.[2]

Pre dokončenie testu potrebujeme vyjadriť výpočet p-hodnoty. Pretože o zamietnutí H_0 vypovedajú hodnoty kritéria F-pomer oveľa väčšie než 1, p-hodnotu vyjadríme ako:

$$p - hodnota = 1 - F_0(X_{OBS}),$$

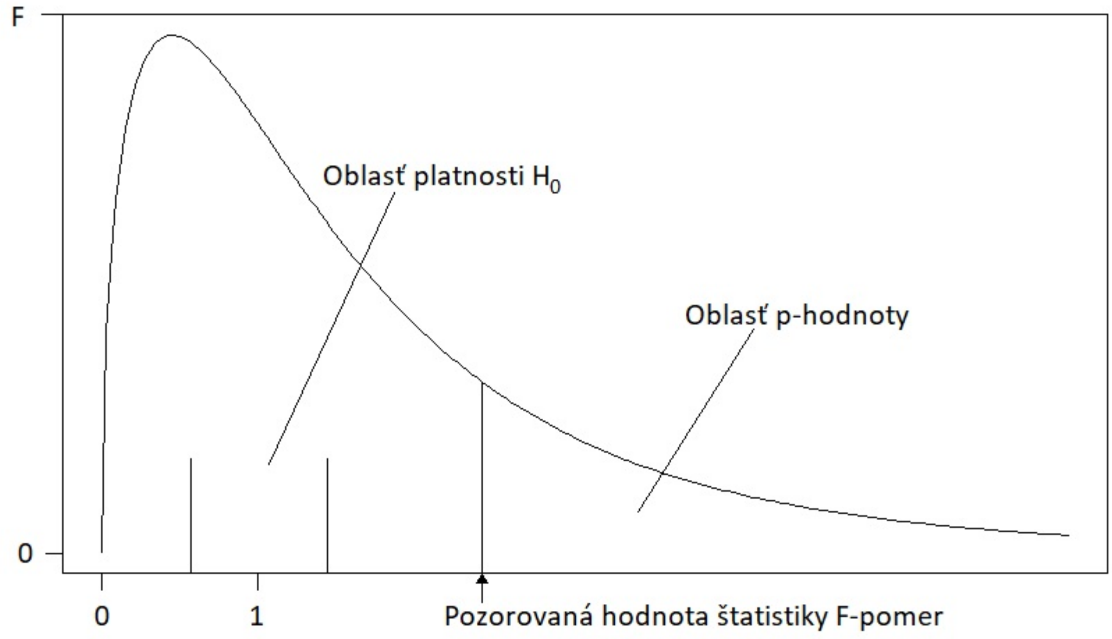
kde $F_0(X)$ je distribučná funkcia Fisher-Snedecorovho rozdelenia s k-1 stupňami voľnosti v čitateli a s N-k stupňami voľnosti v menovateli. Nasledujúci obrázok zobrazuje použitie F-pomeru pri rozhodovaní o pravdivosti H_0.

2.4 Tabuľka ANOVA

Výsledky výpočtov sa pre prehľadnosť zapisujú do tzv. tabuľky jednofaktorovej analýzy rozptylu alebo tabuľky ANOVA.[1]

Zdroj variability	Počet stupňov voľnosti	Súčet štvorcov	Výberový rozptyl	F-pomer	p-hodnota
Medzitriedna	$k-1$	SS_B	S_B^2	$F = \frac{S_B^2}{S_W^2}$	$1 - F_0(X_{OBS})$
Vnútorná	$N-k$	SS_W	S_W^2	–	–
Totálna	$N-1$	SS_T	–	–	–

Tabuľka 1: Tabuľka ANOVA



Obr. 2: P-hodnota pre testovaciu štatistiku F-pomer

Pre lepšie pochopenie si problematiku ukážeme na príklade.

Príklad: Firma vyrábajúca lieky chce na trh uviesť nové antidepresívum. Na overenie efektu liekov spustili výskum, kde bolo vybratých 15 pacientov s klinickou depresiou, ktorí boli náhodne rozdelení do jednej z troch skupín: Skupina 1: Placebo efekt, Skupina 2: Malá dávka liekov, Skupina 3: Normálna dávka liekov. Po 4 týždňoch danej liečby podstúpili pacienti test, tzv. "Beck Depression Inventory". Test sa skladá z 21 otázok s odpoveďami s viacerými možnosťami a čím vyššie skóre pacient dosiahne, tým väčšou depresiou trpí.
Výsledky testov sú zobrazené v nasledujúcej tabuľke:

Poznámka 2.4.1: Typ liečby je v našom prípade varianta nominálnej premennej A, čiže samotná liečba je náš *faktor*, ktorý od seba rozlišuje pacientov. Stupeň depresie je potom X.

Riešenie: Budeme teda skúmať závislosť stupňa depresie (počet bodov z testu) na podávanej liečbe. Položíme nulovú hypotézu:

$$H_0 : \mu_{placebo} = \mu_{malá} = \mu_{normálna}$$

čiže stredné hodnoty počtu bodov v jednotlivých skupinách sa štatisticky významne nelíšia.

Placebo efekt	Malá dávka lieku	Normálna dávka lieku
23	23	17
24	17	16
20	19	17
21	21	15
19	20	18

Tabuľka 2: Výsledky testu "Beck Depression Inventory"po 4 týždňoch liečby

Oproti alternatívnej hypotéze, ktorá hovorí, že sa aspoň jedna dvojica stredných hodnôt štatisticky významne líši (antidepresívum má pozitívne alebo negatívne účinky na pacientov trpiacich depresiou):

$$H_A : \neg H_0$$

Poznámka 2.4.2: Pred použitím testu ANOVA musíme vždy overiť jej predpoklady, tzn. nezávislosť výberu (účastnikmi štúdie bolo 15 nezávisle vybraných pacientov trpiacich depresiou), zhodnosť rozptylov alebo homoskedasticitu (do každej skupiny bolo zaradených 5 pacientov) a normalitu výberov (dáta pochádzajú z normálneho rozdelenia, čo je možné overiť Shapiro–Wilkovým testom pre každý výber samostatne). Predpokladom analýzy rozptylu sa budeme bližšie venovať v 3. kapitole.

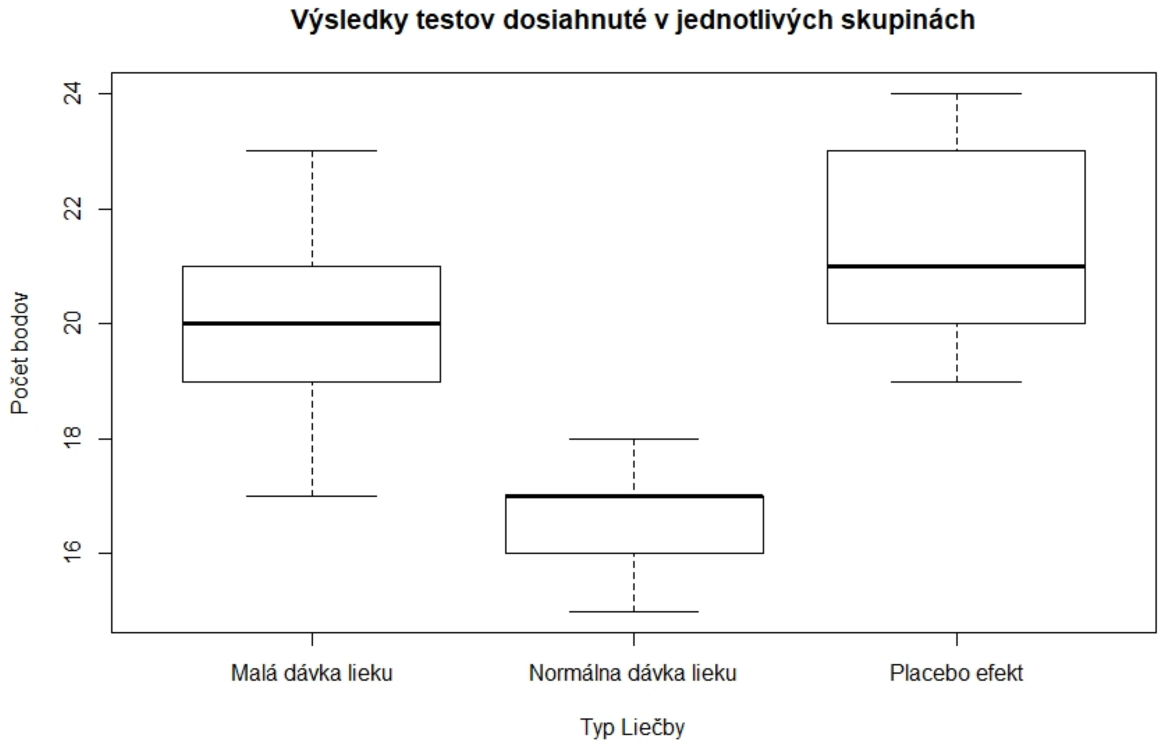
Začneme vizualizáciou dát pomocou krabicového grafu.

Z grafu vidíme, že v dátach nie sú obsiahnuté žiadne odľahlé pozorovania, ktoré by nám mohli znehodnotiť ďalšie testovanie. Vypočítame potrebné hodnoty a dosadíme do tabuľky ANOVA.

Zdroj variability	Počet stupňov voľnosti	Súčet štvorcov	Výberový rozptyl	F-pomer	p-hodnota
Medzitriedna	2	$60,93$	$3,53$	$0,16$	0.0048
Vnútorná	12	$42,40$	$30,47$	–	–
Totálna	14	$103,33$	–	–	–

Tabuľka 3: Výsledky výpočtov dosadené do tabuľky ANOVA

Na hladine významnosti 0,05 nemôžeme potvrdiť nulovú hypotézu, a teda stredné hodnoty jednotlivých výberov sú štatisticky významne odlišné.



Obr. 3: Viacnásobný krabicový graf výsledkov testov v jednotlivých skupinách

2.5 Post Hoc analýza

V prípade, že by dôsledkom nášho testu ANOVA bolo vyvrátenie nulovej hypotézy, mohli by sme prehlásiť, že aspoň dva populačné výberové priemery sa štatisticky významne líšia. Z doterajšej analýzy by sme však nevedeli určiť, pre ktorú dvojicu výberov platí toto tvrdenie. Preto v prípade zamietnutia nulovej hypotézy robíme ďalší krok, ktorý sa nazýva post hoc analýza. Post hoc analýza spočíva v porovnávaní výberových priemerov všetkých dvojíc populácii.

Pre tieto viacnásobné pozorovania existuje niekoľko metód. Medzi najznámejšie a najjednoduchšie patrí tzv. LSD-metóda z angl. *Least Significant Difference.* Spočíva v aplikácii dvojvýberového t-testu pre každý pár výberových priemerov. Použijeme však upravený t-test, ktorý je založený na LSD štatistike:

$$(LSD)_{i,j} = \frac{\bar{X}_i - \bar{X}_j}{S_W\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}} -> t_{N-k},$$

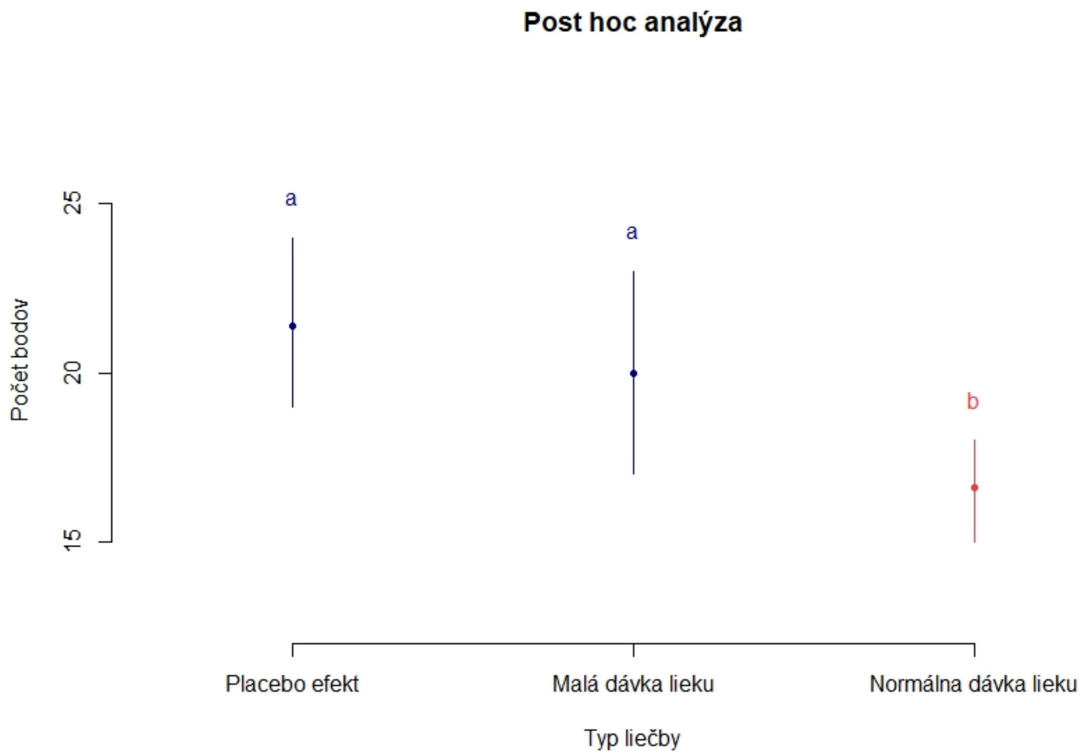
kde

$$S_W = \sqrt{S_W^2} = \sqrt{\frac{SS_W}{N-k}}.$$

Táto štatistika má Studentovo rozdelenie s $N - k$ stupňami volnosti.

Pokračujme v našom príklade o nových vyvinutých liekoch na depresiu. Keďže sme zamietli nulovú hypotézu, potrebujeme pre lepšiu predstavu urobiť post hoc analýzu. V programovacom jazyku R sme spustili LSD metódu a dostávame graf 4.

V tabuľke 4 a na obrázku číslo 4 môžeme vidieť, že medzi Skupinou 1 a Skupinou 2 nie je štatisticky významný rozdiel, boli testom teda priradené do spoločnej skupiny a (p-honota>0,05). Malá dávka lieku teda významne nepomáha pacientom s potlačovaním účinkov depresie. Avšak keď sa pozrieme na testom označenú skupinu b, v ktorej sa nachádza skupina pacientov, ktorým bola podávaná vyššia dávka lieku, môžeme vidieť, že rozdiel oproti výsledkom pacientov, ktorým boli podávane malé dávky lieku alebo dávka vo forme placeba, je štatisticky významný. Taktiež vidíme, že priemerný počet bodov v Skupine 3 (vyššia dávka lieku) je menší ako v ostatných dvoch skupinách, a teda nové vyvíjané antidepresívum môžme považovať za účinné v boji proti depresii pri vyšších dávkach.



Obr. 4: Graficky znázornené rozdiely v stredných hodnotách jednotlivých testovacích skupín

Dvojica výberov	p-hodnota	Významný rozdiel
Placebo efekt - Malá dávka lieku	0,4879	Nie
Malá dávka lieku - Normálna dávka lieku	0,0356	Áno
Placebo efekt - Normálna dávka lieku	0,0043	Áno

Tabuľka 4: p-hodnoty porovnávaných stredných hodnôt dvojíc výberov rozdelených podľa typu liečby

3 ANOVA pri porušení predpokladov

Analýza rozptylu bola pôvodne navrhnutá pre vyvážené triedenie. Pod spojením vyvážené triedenie si predstavujeme jednotlivé použité výbery s rovnakým rozsahom. Tento predpoklad je v praxi len málokedy splnený. Platí však, že čím podobnejšie sú rozsahy jednotlivých výberov, tým sú výsledky testu vierohodnejšie. [2].
ANOVA ako parametrický test predpokladá:

- Nezávislosť výberu
- Normalita rozdelenia
- Homoskedasticita

O jednotlivých predpokladoch a o ďalšom postupe pri ich nesplnení, si povieme v nasledujúcich podkapitolách. V krátkosti si rozoberieme aj jednotlivé alternatívne testy pri nesplnení predpokladov.

3.1 Nezávislosť výberu

Nezávislosť výberu je dôležitým predpokladom. V prípade jeho nesplnenia by sme mohli dostať nespoľahlivé výsledky. Tento predpoklad je možné testovať napr. Friedmanovým testom.

3.1.1 Friedmanov test

Definíciu preberieme zo zdroja [4].

Nech X_{ij} sú nezávislé náhodné veličiny so spojitými distribučnými funkciami F_{ij} pre $i = 1, ..., m$ a $j = 1, ..., k$. Friedmanovým testom sa testuje hypotéza H_0, že F_{ij} nezávisí na j.

Pre každé i zvlášť sa určí poradie veličiny X_{ij}. Ide teda o určenie poradia medzi veličenami $X_{i1}, ..., X_{ik}$. Za platnosti H_0 je splnené podmienka:

H_0': pre každé i je vektor $(R_{i1}, ..., R_{ik})$' rovný ktorejkoľvek permutácii čísel $1, ..., k$ s rovnakou pravdepodobnosťou 1/k! a všetky vektory $(R_{i1}, ..., R_{ik})$' pre $i = 1, ..., k$ sú na sebe nezávislé.

Súčet poradia i-tého výberu označíme ako:

$$R_j = \sum_{i=1}^{m} R_{ij}$$

Platí:

$$Q = -3m(k+1) + \frac{12}{mk(k+1)} \sum_{j=1}^{k} R_j^2 \tag{19}$$

Výpočty sa robia podľa vzorca 19. Hypotézu H_0' (a teda aj hypotézu H_0) zamietame, keď Q prekročí kritickú hodnotu na hladine α. Pri väčších hodnotách m sa za túto hodnotu berie $\chi^2_{k-1}(\alpha)$.

3.2 Normalita rozdelenia

Na porušenie podmienky normality rozdelenia ANOVA nie je príliš citlivá, najmä ak majú jednotlivé výbery rozsah nad 30. Pri výraznejšom porušení normality rozdelenia sa odporúča použit Kruskal-Wallisov test, ktorý je neparametrickou obdobou analýzy rozptylu.
Na overenie tohto predpokladu je možné použiť Shapiro-Wilkov test normality. Jeho nevýhodou je však menšia citlivosť.

3.3 Homoskedasticita

Mierne porušenie predpokladu homoskedasticity, inak aj homogenity rozptylov, testu ANOVA taktiež nevadí, avšak pri väčšom porušení sa, rovnako ako pri väčšom porušení normality rozdelenia, odporúča Kruskal-Wallisov test.

Na overenie homoskedasticity máme na výber hneď z niekoľkých testov. Je možné použiť Bartlettov test, Levenov test alebo Cochranov test.

Testujeme teda nulovú hypotézu:

$$H_0 : \sigma_1^2 = ... = \sigma_k^2, \tag{20}$$

kde k je počet výberov. Oproti alternatívnej:

$$H_A : \neg H_0$$

3.3.1 Bartlettov test

Testovým kritériom u Bartlettovho testu[6] je veličina B, ktorá definovaná ako

$$B = \frac{1}{C}[(N-k)]lnMS_E - \sum_{i=1}^{k}(n_i - 1)lns_i^2],$$

kde

$$MS_E = \frac{1}{N-k}\sum_{i=1}^{k}(n_i - 1)s_i^2$$

MS_E nazývame reziduálny rozptyl.

$$C = 1 + (\sum_{i=1}^{k}\frac{1}{n_i - 1} - \frac{1}{N-k})\frac{1}{\alpha(k-1)}$$

Za platnosti H_0 má testovacia štatistika približne χ^2 rozdelenie s $N-k$ stupňami voľnosti. Potom

$$p - hodnota = 1 - F_0(x_{OBS}),$$

kde $F_0(x)$ je distribučná funkcia χ^2 s $N-k$ stupňami voľnosti.

Bartlettov test je citlivý na porušenie predpokladov normality dát. V prípade porušenia predpokladu volíme Levenov test.

3.3.2 Levenov test

Levenov test je síce menej citlivý na porušenie predpokladu normality rozdelenia, avšak má slabšiu silu testu.[6]

Nech $Z_{ij} = |X_{ij} - \bar{X}_i|$. Označíme:

- $$\bar{Z}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Z_{ij}$$

- $$\overline{\bar{Z}_i} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} Z_{ij}$$

- $$SS_{ZW} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Z_{ij} - \bar{Z})^2$$

- $$SS_{ZB} = \sum_{i=1}^{k} n_i(\bar{Z}_i - \overline{\bar{Z}})^2$$

V prípade, že môžeme prehlásiť nulovú hypotézu za platnú, potom štatistika:

$$F_{ZB} = \frac{\frac{SS_{ZB}}{k-1}}{\frac{SS_{ZW}}{N-k}}$$

má približne Fisher-Snedecorovo rozdelenie s $k-1$ stupňami volnosti v čitateli a $N-k$ stupňami volnosti v menovateli. Potom

$$p - hodnota = 1 - F_0(x_{OBS}),$$

kde $F_0(x)$ je distribučná funkcia Fisher-Snedecorovho rozdelenia s $k-1$ stupňami volnosti v čitateli a $N-k$ stupňami volnosti v menovateli.

Levenov je test je založený na analýze rozptylu absolutných hodnôt centrovaných pozorovaní.

3.3.3 Cochranov test

Testovacím kritériom je veličina G, ktorá je definovaná ako [2]

$$G = \frac{S^2_{max}}{S^2_1 + ... + S^2_I}$$

K zamietnutiu nulovej hypotézy vedú vysoké pozorované hodnoty G_{max}.[2]

3.4 Kruskal-Wallisov test

Ako sme už spomínali, Kruskal-Wallisov test je neparametrickou obdobou analýzy rozptylu. Netestuje zhodu konkrétnych parametrov, ale zhodu výberových distribučných funkcií porovnávaných súborov s tým, že kľúčovým predpokladom je nezávislosť pozorovaných hodnôt.
Majme k porovnávaných výberov, potom nulovú hypotézu Kruskal-Wallisovho testu vyjadríme ako
$H_0 : F_1(\mathrm{x}) = F_2(\mathrm{x}) = ... = F_k(\mathrm{x})$
oproti alternatívnej
H_A: Neplatí H_0

Hlavnou myšlienkou Kruskal-Wallisovho testu je, že za platnosti H_0 sú zlučené hodnoty zo všetkých výberových súborov tak dobre premiešané, že priemerné poradia odpovedajúce jednotlivým súborom sú podobné. Pre výpočet testu teda opäť zoradíme všetky pozorovania podľa veľkosti (ako keby pochádzali z jedného výberu) a priradíme jednotlivým hodnotám poradia (R_{ij} bude označovať poradie j-tej hodnoty v i-tej skupine).
Označme k celkový počet skupín, n celkový počet pozorovaní a n_1,n_2,...,n_k počty pozorovaní v

jednotlivých skupinách n = n_1,n_2,...,n_k. ďalej označme T_i súčet poradia v i-tej skupine:

$$T_i = \sum_{j=1}^{n_i} R_{ij}$$

Potom testová štatistika Kruskal-Wallisovho testu bude mať tvar

$$Q = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{T_i^2}{n_i} - 3(n+1)$$

Je možné ukázať, že testová štatistika Q má za platnosti nulovej hypotézy chí-kvadrát rozdelenia pravdepodobnosti s parametrom k-1.
Nulovú hypotézu H_0 potom zamietame na hladine významnosti α, keď je realizácia testovej štatistiky Q väčšia než kritická hodnota príslušná hladine významnosti α.
Pre malé veľkosti súboru je potrebné porovnať štatistiku Q s tabuľkami pre Kruskal-Wallisov test.

Príklad. Výrobca koláčov v prášku má 4 nové recepty a chce zistiť, či sa ich kvalita líši. Upiekol preto 5 koláčov z každého druhu a dal ich ohodnotiť porote.

Recept A	72	88	70	87	71
Recept B	85	89	86	82	88
Recept C	94	94	88	87	89
Recept D	91	93	92	95	94

Tabuľka 5: Ohodnotenie koláčov porotou

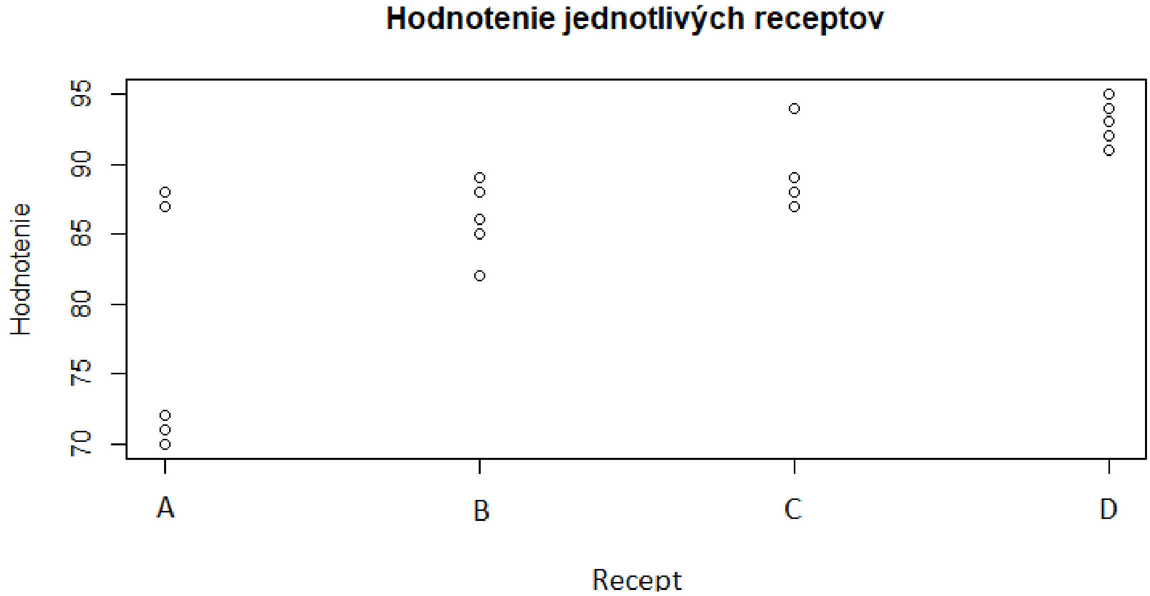
Na hladine významosti 0,05 testujte hypotézu, že sa ich kvalita nelíši.
Riešenie. Položíme nulovú a alternatívnu hypotézu:

H_0: Kvalita koláčov sa nelíši

oproti

H_A: Kvalita koláčov sa líši

Graficky znázornime počet bodov pre jednotlivé recepty pomocou bodového grafu:



Obr. 5: Počet bodov pre jednotlivé recepty

Na obrázku 5 vidíme grafické rozloženie bodov pre recepty. Použili sme bodový graf kvôli malému počtu dát. Pokračujeme spustením samotného Kruskal-Wallisovho testu v softwari RStudio. Výsledná $p - hodnota = 0.005737$, pričom $0.005737 < 0,05$. Dôsledkom je zamietnutie nulovej hypotézy na hladine významnosti 0,05. Môžeme potom povedať, že kvalita koláčov podľa receptov A, B, C a D sa štatisticky významne líši.

Na záver určíme, ktoré dvojice výberov majú významné rozdiely. Použijeme porovnanie všetkých dvojíc výberov pomocou Mann-Whitney U-testu a výsledky všetkých porovnaní zapíšeme do prehľadnej tabuľky. Nerobíme teda iba jednu post hoc analýzu, ale niekoľko, v závislosti na zadanom rozdelení podľa ďalšej premennej.

Dvojica výberov	p-hodnota	Štatisticky významny rozdiel
A-B	$0,138 > 0,05$	Nie
A-C	$0,066 > 0,05$	Nie
A-D	$0,043 < 0,05$	Áno
B-C	$0,042 < 0,05$	Áno
B-D	$0,039 < 0,05$	Áno
C-D	$0,225 > 0,05$	Nie

Tabuľka 6: p-hodnoty pri porovnávaní stredných hodnôt dvojíc výberov podľa receptov

Z výsledkov v tabuľke 6 nemôžeme zamietnúť štatisticky významný rozdiel v kvalite koláčov medzi dvojicami receptov A-D, B-C a B-D.

4 Porovnanie testu ANOVA s odpovedajúcimi dvojvýberovými testami

4.1 Dvojvýberový t-test

Jeden z najpoužívanješích testov, ktorý umožnuje porovnávať dve populácie na základe dvoch nezávislých výberov, je dvojvýberový t-test. [3] Nezávislosť výberov sa v praxi zaisťuje tým, že výbery obsahujú rozdielne prvky. Ide o test o zhode dvoch stredných hodnôt. Test je založený na predpoklade, že výbery pochádzaju z normalného rozdelenia.

Majme dva nezávislé výbery z dvoch normálne rozdelených populácii, z ktorých porovnávame stredné hodnoty:

$Výber\ 1 : (X_1, X_2, ..., X_{n_1})$
$Výber\ 2 : (Y_1, Y_2, ..., Y_{n_2})$

Pochádzajú teda z rozdelenia $N(\mu_1; \sigma_1^2)$, resp. $N(\mu_2; \sigma_2^2)$.

Budeme používať značenie:

- Výberové priemery jednotlivých výberov:

$$\bar{X} = \frac{\sum_{i=1}^{m} X_i}{n_1}, \tag{21}$$

resp.

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n_2}. \tag{22}$$

- Výberový rozptyly jednotlivých výberov:

$$\bar{S_1^2} = \frac{\sum_{i=1}^{m} (X_i - \bar{X})^2}{n_1 - 1}, \tag{23}$$

resp.

$$\bar{S_2^2} = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n_2 - 1} \tag{24}$$

Aj pre dvojvýberový test si sformulujeme problém. Testujeme nulovú hypotézu, ktorá hovorí, že stredné hodnoty jednotlivých výberov sa štatisticky významne nelíšia:

$$H_0 :\ \mu_1 = \mu_2$$

Čo sa týka alternatívnej hypotézy H_A, máme na výber z troch možností:

$$H_A :\ \mu_1 < \mu_2\ alebo\ \mu_1 - \mu_2 < 0 \tag{25}$$

$$H_A: \; \mu_1 > \mu_2 \, alebo \, \mu_1 - \mu_2 > 0 \tag{26}$$

$$H_A: \; \mu_1 \neq \mu_2 \, alebo \, \mu_1 - \mu_2 \neq 0 \tag{27}$$

Voľba vhodnej alternatívnej hypotézy tentokrát závisi na priemeroch jednotlivých výberov. V prípade, ak $\bar{X}$ je výrazne menšia ako $\bar{Y}$, volíme alternatívu 25. Ak je $\bar{X}$ výrazne väčšia ako $\bar{Y}$, volíme 26. Ak sa hodnota $\bar{X}$ pohybuje v blízkosti $\bar{Y}$, volíme alternatívu v tvare 27.

Pri voľbe testovacieho kritéria musíme zohľadiť informácie, ktoré máme o jednotlivých smerodajných odchýlkach σ_1 σ_2. Keďže túto prácu venujeme analýze rozptylu a túto kapitolu dvojvýberovému testu odpovedajúcemu testu ANOVA, spomenieme len prípad, kedy odchýlky nepoznáme, ale predpokladáme, že su zhodné.

Volíme potom dvojvýberový t-test. Za testové kritérium volíme štatistiku:

$$T(\underline{X}) = T_2 = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} -> t_{n_1+n_2-2}, \tag{28}$$

kde

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \tag{29}$$

4.2 Porovnanie dvojvýberového t-testu s testom ANOVA

V predošlých kapitolách sme si povedali, že pri testovaní zhody stredných hodnôt k>=2 výberov a pri splnení predpokladov použijeme analýzy rozptylu. V rovnakom prípadem avšak len pri dvoch takýchto výberoch použijeme dvojvýberový t-test. Tieto tvrdenia nám napovedajú, že analýzu rozptylu použitú pre dva nezávislé a náhodné výbery, pochádzajúce z normálneho rozdelenia so zhodným rozptylom, môžeme považovať za ekvivalentnú k dvojvýberovému t-testu pre stredné hodnoty s neznámymi, avšak predpokladane zhodnými rozptylmi. V tejto podkapitole si overíme správnosť tohoto výroku.

Príklad. 20 študentov bolo náhodne rozdelených do rovnako veľký skupín. V skupine 1 bola študentom počas testu pustená hudba na určitej úrovni hlasitosti. V skupine 2 písali študenti test bez hudby. Cieľom pokusu bolo zistiť, ktorá skupina študentov dosiahne lepšie výsledky, a teda či hudba ovplyvňuje dosiahnuté skóre z testov. Porovnajte stredné hodnoty dosiahnutného skóre v jednotlivých skupinách.

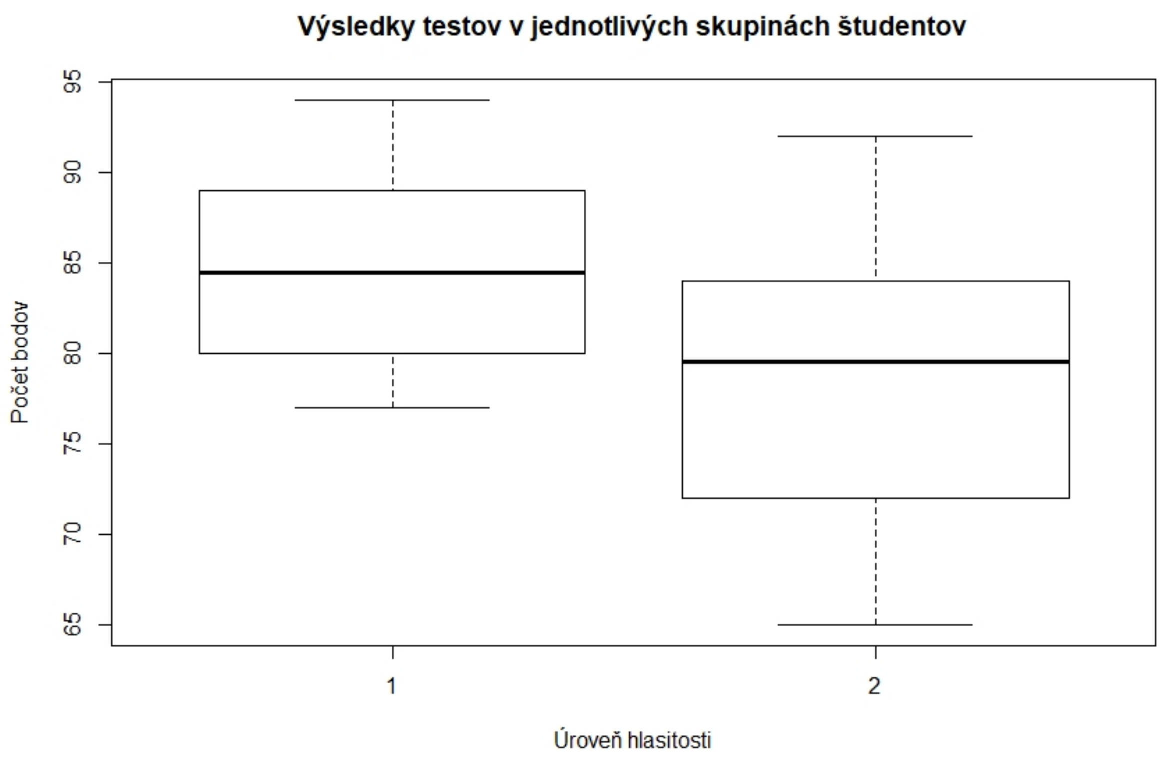
Riešenie. Príklad vyriešime pomocou dvojvýberového testu o stredných hodnotách a pomocou testu ANOVA.

Výsledky testov sú zobrazené v nasledujúcej tabuľke:

Skupina 1	79	84	65	70	79	83	72	92	80	89
Skupina 2	84	80	92	78	88	85	94	77	82	89

Tabuľka 7: Výsledky testov študentov v jednotlivých skupinách

Výsledky testov v závislosti na úrovni hlasitosti hudby si zobrazíme krabicovým grafom. Na grafe vidíme, že výbery neobsahujú žiadne odľahlé pozorovania, ktoré by mohli znehodnotiť ďalšie testovanie.



Obr. 6: Graf výsledkov testov študentov v skupine 1 a v skupine 2

Na overenie normality dát sme použili Shapiro-Wilkov test. Položili sme nulovú a alternatívnu hypotézu pre jednotlivé výbery:

$$H_0 : \text{dáta pochádzajú z normálneho rozdelenia}$$

$$H_A : \text{dáta nepochádzajú z normálneho rozdelenia}$$

Pre výber študentov zo skupiny 1 (bez hudby) sme získali $p - hodnotu = 0,8016$. Pre výber študentov zo skupiny 2 (s hudbou) sme získali $p - hodnotu = 0,2394$. Na hladine významnosti 0,05 sme nezamietli H_0, keďže obe $p - hodnoty > 0,05$. Z toho vyplýva, že môžeme predpokladať normalitu dát.

Teraz môžeme pristúpiť k samotným testom. Na vyhodnotenie testov použijeme programovací jazyk R. Najprv však určíme nulovú a alternatívnu hypotézu:

$$H_0 : \mu_1 = \mu_2$$
$$H_A : \mu_1 \neq \mu_2$$

```
        Two Sample t-test

data:  Body by Hlasitosť
t = 1.7337, df = 18, p-value = 0.1001
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.186119 12.386119
sample estimates:
mean in group 1 mean in group 2
           84.9            79.3
```

Obr. 7: Výsledky dvojvýberového t-testu o zhodných stredných hodnotách v RStudiu

```
            Df Sum Sq Mean Sq F value Pr(>F)
Hlasitosť    1  156.8  156.80   3.006    0.1
Residuals   18  939.0   52.17
```

Obr. 8: Výsledky testu ANOVA v RStudiu

Na obrázkoch 7 a 8 môžeme vidieť, že $p - hodnota_{t-test} = p - hodnota_{ANOVA}$. Mohli by sme namietať, že u výsledkov z testu ANOVA má p-hodnota len jedno desatinné miesto, dôvodom je, že p-hodnota bola zaokrúhlená na tretie desatinné miesto, čiže na 0,100.

Keďže $p - hodnota > 0,05$, nezamietame nulovú hypotézu. V stredných hodnotách výsledkov z testu, ktorý podstúpili študenti oboch skupín, neexistuje štatisticky významný rozdiel, a teda hudba neovplyvňuje študentov pri písaní testu.

Zhodné p-hodnoty však indikujú, že dvojvýberový t-test a test ANOVA sú ekvivalentné. Všimnime si hodnoty t u dvojvýberového t-testu a hodnoty F u testu ANOVA. Je zjavné, že $1,7337^2 = 3,006$, takže $t^2(n_1 + n_2 - 2) = F(1, n_1 + n_2 - 2)$. V nasledujúcich krokoch sa pokúsime dokázať toto tvrdenie.

Veta 4.2.1. $F(1, n_1 + n_2 - 2) = t^2(n_1 + n_2 - 2)$

Dôkaz. Z kapitoly 2 vieme, že F-pomer je definovaný vzťahom 18 a pri platnosti H_0 má Fisher-Snedecorove rozdelenie s k-1 stupňami voľnosti v čitateli a s N-k stupňami voľnosti v menovateli $F_{(k-1,N-k)}$, kde k je počet výberov a N je celkový počet pozorovaní v obidvoch výberoch. Čo sa týka dvojvýberového t testu využijeme vzťah 28 a 29.

Hovoríme o dvoch nezávislých náhodných výberoch, pochádzajúcich z normálneho rozdelenia, môžeme teda zapísať $F(2 - 1, n_1 + n_2 - 2)$.

- Položme nulovú a alternatívnu hypotézu.

$$H_0 : \mu_1 = \mu_2$$
$$H_A : \mu_1 \neq \mu_2$$

- Najprv upravíme vzťah 28.

$$T(\underline{X}) = \frac{(\bar{X_1} - \bar{X_2}) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

kde $\mu_1 - \mu_2 = 0$, čo vyplýva z nulovej hypotézy. Dosadíme s_p zo vzorca 29:

$$T(\underline{X}) = \frac{(\bar{X_1} - \bar{X_2})}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

Pomocou základných algebraických vzťahov urobíme ďalšie úpravy a dostávame:

$$T(\underline{X}) = \frac{(\bar{X_1} - \bar{X_2})}{\sqrt{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}}.\sqrt{\frac{n_1 n_2}{n_1 + n_2}.(n_1 + n_2 - 2)}$$

Položme

$$t(n_1 + n_2 - 2) = \frac{(\bar{X_1} - \bar{X_2})}{\sqrt{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}}.\sqrt{\frac{n_1 n_2}{n_1 + n_2}.(n_1 + n_2 - 2)} \tag{30}$$

- Teraz postupne upravíme vzťah 18.

$$F_{(1,n_1+n_2-2)} = \frac{SS_B}{SS_W}(n_1 + n_2 - 2)$$

Dosadením zo vzorcov 8 a 9 dostávame:

$$F_{(1,n_1+n_2-2)} = \frac{n_1(\bar{X_1} - \bar{X})^2 + n_2(\bar{X_2} - \bar{X})^2}{(n_1-1)S_1^2 + (n_2-1)S_2^2}(n_1+n_2-2)$$

Položíme

$$k = \frac{(n_1+n_2-2)}{(n_1-1)S_1^2 + (n_2-1)S_2^2}.$$

Vyjadríme si $\bar{X}$, čiže priemer zo všetkých pozorovaných hodnôt ako:

$$\bar{X} = \frac{n_1\bar{X_1} + n_2\bar{X_2}}{n_1+n_2},$$

kde $\bar{X_1}$ a $\bar{X_2}$ sú priemery jednotlivých výberov.
Z toho vyplýva:

$$F_{(1,n_1+n_2-2)} = k[n_1(\bar{X_1} - \frac{n_1\bar{X_1} - n_2\bar{X_2}}{n_1+n_2})^2 + n_2(\bar{X_2} - \frac{n_1\bar{X_1} - n_2\bar{X_2}}{n_1+n_2})^2]$$

Ďalej pokračujeme základnými úpravami zlomkov:

$$F_{(1,n_1+n_2-2)} = k[n_1n_2^2\frac{(\bar{X_1} - \bar{X_2})^2}{(n_1+n_2)^2} + n_2n_1^2\frac{(\bar{X_1} - \bar{X_2})^2}{(n_1+n_2)^2}] = k\frac{(\bar{X_1} - \bar{X_2})^2 n_1n_2}{(n_2+n_1)}$$

Dosadíme za k:

$$F(1, n_1+n_2-2) = \frac{(n_1+n_2-2)n_1n_2}{(n_1-1)S_1^2 + (n_2-1)S_2^2}\frac{(\bar{X_1} - \bar{X_2})^2}{(n_2+n_1)} \tag{31}$$

Z výrazov 31 a 30 vidíme, že rovnosť platí pre:

$$F(1, n_1+n_2-2) = t^2(n_1+n_2-2)$$

Týmto sme dokázali trvdenie, že jednofaktorová analýza rozptylu a dvojvýberový t-test su ekvivalentné. Produkujú teda zhodné rozhodnutie o platnosti nulovej hypotézy a ich p-hodnoty sú zhodné. Avšak táto rovnosť platí iba v prípade pri dvoch nezávislých náhodných výberoch pochádzajúcich z normálneho rozdelenia.

Na tomto mieste sa ponúka otázka. Je možné, že by test ANOVA a dvojvýberový t-test boli ekvivalentnými aj pre viac ako dva výbery?

Povedzme, že máme n náhodných nezávislých výberov a chcem porovnať ich stredné hodnoty pomocou dvojvýberového t-testu. Mohlo by sa zdať, že by stačilo utvoriť dvojice náhodných výberov a na všetky aplikovať dvojvýberový t-test. Týchto testov by potom bolo $\binom{2}{n} = \frac{n(n-1)}{2}$.

Keby sme každý test urobili na hladine významnosti α, bola by výsledná hladina významnosti testu omnoho vyššia ako α. Týmto by sme celý test znehodnotili. Tieto dva testy sú ekvivaletné teda naozaj len pre dva nezávislé náhodné výbery.

5 Záver

Cieľom tejto práce bolo popísať analýzu rozptylu a jej vlastnosti ako parametrického testu. Ukázali sme si postup testu pri splnení predpokladov, potvrdení, ale aj zamietnutí nulovej hypotézy.

Taktiež sme sa zaoberali analýzou medzitriedneho výberového rozptylu, jednotlivými postupmi v prípade nesplnenia predpokladov a porovnávaniu analýzy rozptylu s odpovedajúcimi dvojvýberovými testami.

Analýzou medzitriedneho rozptylu sme dokázali, že ho pri potvrdení platnosti nulovej hypotézy môžeme považovať za nestranný odhad rozptylu.

Čo sa týka porovnávania testu ANOVA a dvojvýberového t-testu, na základe rozdelení $t(n_1 + n_-2)$ a $F(n_1 + n_-2)$ sme dokázali, že v prípade dvoch nezávislých výberov môžeme tieto testy považovať za ekvivalentné. Dostávame zhodnú výslednú p-hodnotu.

V prípade nesplnenia predpokladov analýzy rozptylu musíme použiť inú voľbu testov, aby nedošlo k znehodnoteniu analýzy. Predpoklad nezávislosti môžeme overiť Friedmanovým testom. Pri porušení normality, tak ako aj pri porušení homoskedasticity použijeme Kruskal-Wallisov test, ktorý je neparametrickou obdobou testu ANOVA. Predpoklad normality je možné overiť Shapiro-Wilkovým testom normality a predpoklad homoskedascity napr. Bartlettovým, Levenovým alebo Cohranovým testom.

Literatúra

[1] Briš R., Litschmannová M.,STATISTIKA I. pro kombinované a distanční studium, Elektronické skriptum VŠB TU Ostrava, 2004.

[2] Litschmannová M., Úvod do statistiky, Elektronické skriptum VŠB TU Ostrava, 2011

[3] Weerahandi S.: Exact Statistical Methods for Data Analysis, Springer-Verlag New York, Inc., 1995, ISBN 0-387-94360-9.

[4] ANDĚL, Jiří. Statistické metody. Vyd. 3. Praha: Matfyzpress, 2003. ISBN 80-86732-08-8.

[5] ANDĚL, Jiří. Matematická statistika. Praha: SNTL - Nakladatelství technické literatury, 1978.

[6] BUDÍKOVÁ, Marie, Maria KRÁLOVÁ a Bohumil MAROŠ. Průvodce základními statistickými metodami. Praha: Grada Publishing, 2010. Expert. ISBN 978-80-247-3243-5.